From the book *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*.
By David Easley and Jon Kleinberg. Cambridge University Press, 2010.
Complete preprint on-line at http://www.cs.cornell.edu/home/kleinber/networks-book/

Chapter 18

Power Laws and Rich-Get-Richer Phenomena

18.1 Popularity as a Network Phenomenon

For the past two chapters, we have been studying situations in which a person's behavior or decisions depend on the choices made by other people — either because the person's rewards are dependent on what other people do, or because the choices of other people convey information that is useful in the decision-making process. We've seen that these types of coupled decisions, where behavior is correlated across a population, can lead to outcomes very different from what we find in cases where individuals make independent decisions.

Here we apply this network approach to analyze the general notion of *popularity*. Popularity is a phenomenon characterized by extreme imbalances: while almost everyone goes through life known only to people in their immediate social circles, a few people achieve wider visibility, and a very, very few attain global name recognition. Analogous things could be said of books, movies, or almost anything that commands an audience. How can we quantify these imbalances? Why do they arise? Are they somehow intrinsic to the whole idea of popularity?

We will see that some basic models of network behavior can provide significant insight into these questions. To begin the discussion, we focus on the Web as a concrete domain in which it is possible to measure popularity very accurately. While it may be difficult to estimate the number of people worldwide who have heard of famous individuals such as Barack Obama or Bill Gates, it is easy to take a snapshot of the full Web and simply count

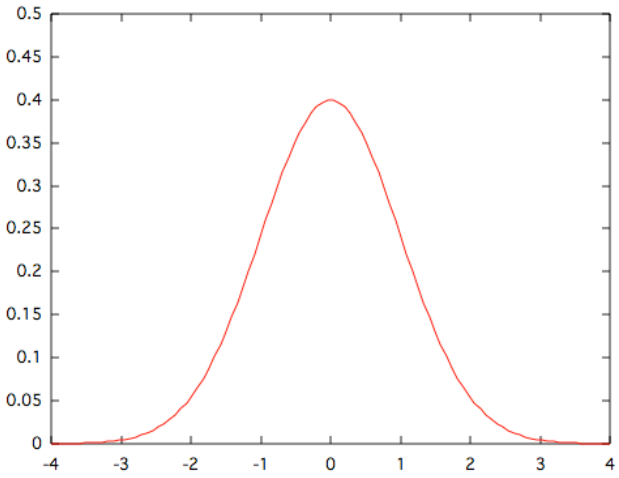

Figure 18.1: The density of values in the normal distribution.

the number of links to high-profile Web sites such as Google, Amazon, or Wikipedia. We will refer to the full set of links pointing to a given Web page as the *in-links* to the page. Thus, we will start by using the number of in-links to a Web page as a measure of the page's popularity; we will keep in mind, however, that this is just one example of a much broader phenomenon.

Early in the Web's history, people had already begun to ask a very basic version of the page popularity question, phrased as follows:

As a function of k, what fraction of pages on the Web have k in-links?

Larger values of k indicate greater popularity, so this is precisely the question of how popularity is distributed over the set of Web pages.

A Simple Hypothesis: The Normal Distribution. Before trying to resolve this question, it's useful to ask what we should expect the answer to be. A natural guess is the *normal*, or *Gaussian*, distribution — the so-called "bell curve" — used widely throughout probability and statistics. While we won't need many details about the normal distribution here, it's worth recalling that it's characterized by two quantities: a mean value, and a standard deviation around this mean. Figure 18.1 shows a plot of the density of values in the normal distribution, scaled so that the mean is 0 and the standard deviation is 1. The basic fact about normal distributions is that the probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially in c.

The normal distribution is a natural guess in our case, since it is ubiquitous across the natural sciences. A result from the early 1900s, the *Central Limit Theorem*, provides a

fundamental explanation for its appearance in so many settings: roughly speaking (and suppressing the full details), the Central Limit Theorem says that if we take any sequence of small independent random quantities, then in the limit their sum (or average) will be distributed according to the normal distribution. In other words, any quantity that can be viewed as the sum of many small independent random effects will be well-approximated by the normal distribution. Thus, for example, if one performs repeated measurements of a fixed physical quantity, and if the variations in the measurements across trials are the cumulative result of many independent sources of error in each trial, then the distribution of measured values should be approximately normal.

How would we apply this in the case of Web pages? If we model the link structure of the Web, for example, by assuming that each page decides independently at random whether to link to any other given page, then the number of in-links to a given page is the sum of many independent random quantities (i.e. the presence or absence of a link from each other page), and hence we'd expect it to be normally distributed. In particular, this would suggest a hypothesis for the answer to our original question: if we believe this model, then the number of pages with k in-links should decrease exponentially in k, as k grows large.

18.2 Power Laws

When people measured the distribution of links on the Web, however, they found something very different. In studies over many different Web snapshots, taken at different points in the Web's history, the recurring finding is that the fraction of Web pages that have k in-links is approximately proportional to $1/k^2$ [80]. (More precisely, the exponent on k is generally a number slightly larger than 2.)

Why is this so different from the normal distribution? The crucial point is that $1/k^2$ decreases much more slowly as k increases, so pages with very large numbers of in-links are much more common than we'd expect with a normal distribution. For example, $1/k^2$ is only one in a million for $k = 1000$, while an exponentially decaying function like 2^{-k} is unimaginably tiny for $k = 1000$. A function that decreases as k to some fixed power, such as $1/k^2$ in the present case, is called a *power law*; when used to measure the fraction of items having value k, it says, qualitatively, that it's possible to see very large values of k.

This provides a quantitative form for one of the points we made initially: popularity seems to exhibit extreme imbalances, with very large values likely to arise. And it accords with our intuition about the Web, where there are certainly a reasonably large number of extremely popular pages. One sees similar power laws arising in measures of popularity in many other domains as well: for example, the fraction of telephone numbers that receive k calls per day is roughly proportional to $1/k^2$; the fraction of books that are bought by k people is roughly proportional to $1/k^3$; the fraction of scientific papers that receive k citations

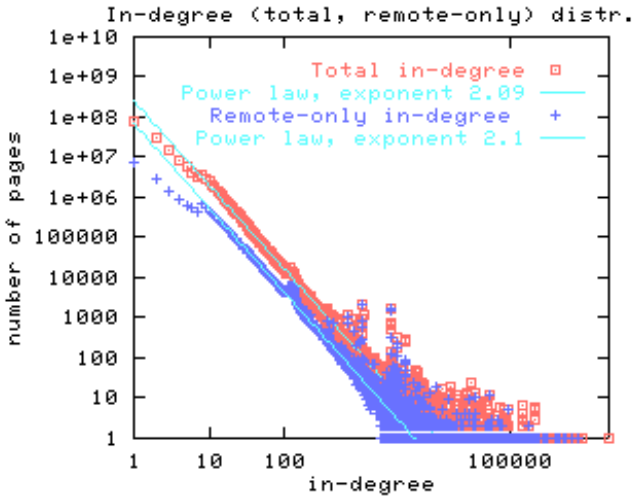

Figure 18.2: A power law distribution (such as this one for the number of Web page in-links, from Broder et al. [80]) shows up as a straight line on a log-log plot.

in total is roughly proportional to $1/k^3$; and there are many related examples [10, 320].

Indeed, just as the normal distribution is widespread in a family of settings in the natural sciences, power laws seem to dominate in cases where the quantity being measured can be viewed as a type of popularity. Hence, if you are handed data of this sort — say, for example, that someone gives you a table showing the number of monthly downloads for each song at a large on-line music site that they're hosting — one of the first things that's worth doing is to test whether it's approximately a power law $1/k^c$ for some c, and if so, to estimate the exponent c.

There's a simple method that provides at least a quick test for whether a dataset exhibits a power-law distribution. Let $f(k)$ be the fraction of items that have value k, and suppose you want to know whether the equation $f(k) = a/k^c$ approximately holds, for some exponent c and constant of proportionality a. Then, if we write this as $f(k) = ak^{-c}$ and take the logarithms of both sides of this equation, we get

$$\log f(k) = \log a - c \log k.$$

This says that if we have a power-law relationship, and we plot $\log f(k)$ as a function of $\log k$, then we should see a straight line: $-c$ will be the slope, and $\log a$ will be the y-intercept. Such a "log-log" plot thus provides a quick way to see if one's data exhibits an approximate power-law: it is easy to see if one has an approximately straight line, and one can read off the exponent from the slope. For example, Figure 18.2 does this for the fraction of Web pages with k in-links [80].

But if we are going to accept that power laws are so widespread, we also need a simple explanation for what is causing them: just as the Central Limit Theorem gave us a very

basic reason to expect the normal distribution, we'd like something comparable for power laws. For example, it's striking how closely the plot in Figure 18.2 follows a straight line for much of the distribution, especially considering how many utterly uncontrollable factors come into play in the formation of the Web's link structure. What underlying process is keeping the line so straight?

18.3 Rich-Get-Richer Models

Ideas from the analysis of information cascades and network effects provide the basis for a very natural mechanism to generate power laws. Just as normal distributions arise from many independent random decisions averaging out, we will find that power laws arise from the feedback introduced by correlated decisions across a population.

It is actually an open and very interesting research question to provide a fully satisfactory model of power laws starting from simple models of individual decision-making (as we did for information cascades). Instead, we will build our model not from the internals of each person's decision-making process, but from the observable consequences of decision-making in the presence of cascades: we will assume simply that people have a tendency to copy the decisions of people who act before them.

Based on this idea, here is a simple model for the creation of links among Web pages [42, 265, 300, 340, 371].

(1) Pages are created in order, and named $1, 2, 3, \ldots, N$.

(2) When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule (which is controlled by a single number p between 0 and 1).

   (a) With probability p, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to this page i.

   (b) With probability $1-p$, page j instead chooses a page i uniformly at random from among all earlier pages, and creates a link *to the page that i points to.*

   (c) This describes the creation of a single link from page j; one can repeat this process to create multiple, independently generated links from page j. (However, to keep things simple, we will suppose that each page creates just one outbound link.)

Part (2b) of this process is the key: after finding a random earlier page i in the population, the author of page j does not link to i, but instead copies the decision made by the author of page i — linking to the same page that i did.

The main result about this model is that if we run it for many pages, the fraction of pages with k in-links will be distributed approximately according to a power law $1/k^c$, where the value of the exponent c depends on the choice of p [68]. This dependence goes in an

intuitively natural direction: as p gets smaller, so that copying becomes more frequent, the exponent c gets smaller as well, making one more likely to see extremely popular pages.

Proving this result would require more intricate analysis than we'll be able to do here, but it is useful to work through some of the informal ideas behind this analysis. First of all, the copying mechanism in (2b) is really an implementation of the following "rich-get-richer" dynamics: when you copy the decision of a random earlier page, the probability that you end up linking to some page ℓ is directly proportional to the total number of pages that currently link to ℓ. Thus, an equivalent way to write our copying process would have been to phrase (2b) as

(2) ...

(b) With probability $1 - p$, page j chooses a page ℓ with probability proportional to ℓ's current number of in-links, and creates a link to ℓ.

Why do we call this a "rich-get-richer" rule? Because the probability that page ℓ experiences an increase in popularity is directly proportional to ℓ's current popularity. This phenomenon is also known as *preferential attachment* [42], in the sense that links are formed "preferentially" to pages that already have high popularity. And the copying model provides an operational story for why popularity should exhibit such rich-get-richer dynamics: essentially, the more well-known someone is, the more likely you are to hear their name come up in conversation, and hence the more likely you are to end up knowing about them as well. The same holds for Web pages, the specific focus of our model here.

The remaining intuition behind the analysis runs as follows. With the rich-get-richer dynamics in place, our model predicts that popularity should grow according to the same rule that governs the growth of bacterial colonies and compound interest: a page's popularity grows at a rate proportional to its current value, and hence exponentially with time. A page that gets a small lead over others will therefore tend to extend this lead; whereas the crux of the Central Limit Theorem is that small independent random values tend to cancel each other out, the rich-get-richer nature of copying actually amplifies the effects of large values, making them even larger. In Section 18.7 at the end of this chapter, we show how to turn this reasoning into a calculation that produces the correct exponent on the power-law distribution.

As with any simple model, the goal is not to capture all the reasons why people create links on the Web, or in any other network, but to show that a simple and very natural principle behind link creation leads directly to power laws — and hence, one should not find them as surprising as they might first appear.

Indeed, rich-get-richer models can suggest a basis for power laws in a wide array of settings, including some that have nothing at all to do with human decision-making. For example, the populations of cities have been observed to follow a power law distribution: the

fraction of cities with population k is roughly $1/k^c$ for some constant c [371]. If we assume that cities are formed at different times, and that, once formed, a city grows in proportion to its current size simply as a result of people having children, then we have almost precisely the same rich-get-richer model — and hence we should not be surprised to see the power law that is in fact present in reality. To take a very different example, researchers in biology have argued (though the data is still too sparse to be sure) that the number of copies of a gene in a genome approximately follows a power-law distribution [99]. If we believe that gene copies arise in large part through mutational events in which a random segment of DNA is accidentally duplicated, then a gene which already has many copies is proportionally more likely to be lying in a random stretch of DNA that gets copied — so "rich" genes (those with many copies) get "richer," and again we should not be surprised to see a power law.

A priori, finding similar laws governing Web page popularity, city populations, and gene copies is quite mysterious; but if one views all these as outcomes of processes exhibiting rich-get-richer effects, then the picture starts to become clearer. Again, one must stress that these are still simple models designed just to approximate what's going on; and there are other classes of simple models designed to capture power-law behavior that we have not discussed here. For example, a parallel thread of research has argued how power laws can arise from systems that are being *optimized* in the presence of constraints [96, 136, 151, 284, 300]. But what all these simple models suggest is that when one sees a power law in data, the possible reasons *why* it's there can often be more important than the simple fact *that* it's there.

18.4 The Unpredictability of Rich-Get-Richer Effects

Given the nature of the feedback effects that produce power laws, it's natural to suspect that for a Web page, a book, a song, or any other object of popular attention, the initial stages of its rise to popularity is a relatively fragile thing. Once any one of these items is well-established, the rich-get-richer dynamics of popularity are likely to push it even higher; but getting this rich-get-richer process ignited in the first place seems like a precarious process, full of potential accidents and near-misses.

This sensitivity to unpredictable initial fluctuations is something that we saw in the previous two chapters as well: information cascades can depend on the outcome of a small number of initial decisions in the population, and a worse technology can win because it reaches a certain critical audience size before its competitors do. The dynamics of popularity suggest that random effects early in the process should play a role here as well. For example, if we could roll time back 15 years, and then run history forward again, would the Harry Potter books again sell hundreds of millions of copies, or would they languish in obscurity while some other works of children's fiction achieved major success? One's intuition suggests the latter. More generally, if history were to be replayed multiple times, it seems likely that

there would be a power-law distribution of popularity each of these times, but it's far from clear that the most popular items would always be the same.

Thought experiments of this type are useful in considering the consequences of our models, but, needless to say, it's difficult to actually implement them as real experiments. Recently, however, Salgankik, Dodds, and Watts performed an experiment that begins to provide some empirical support for this intuition [359]. They created a music download site, populated with 48 obscure songs of varying quality written by actual performing groups. Visitors to the site were presented with a list of the songs and given the opportunity to listen to them. Each visitor was also shown a table listing the current "download count" for each song — the number of times it had been downloaded from the site thus far. At the end of a session, the visitor was given the opportunity to download copies of the songs that he or she liked.

Now, unbeknownst to the visitors, upon arrival they were actually being assigned at random to one of eight "parallel" copies of the site. The parallel copies started out identically, with the same songs and with each song having a download count of zero. However, each parallel copy then evolved differently as users arrived. In a controlled, small-scale setting, then, this experiment provided a way to observe what happens to the popularities of 48 songs when you get to run history forward eight different times. And in fact, it was found that the "market share" of the different songs varied considerably across the different parallel copies, although the best songs never ended up at the bottom and the worst songs never ended up at the top.

Salganik et al. also used this approach to show that, overall, feedback produced greater inequality in outcomes. Specifically, they assigned some users to a ninth version of the site in which no feedback about download counts was provided at all. In this version of the site, there was no direct opportunity for users to contribute to rich-get-richer dynamics, and indeed, there was significantly less variation in the market share of different songs.

There are clear implications for popularity in less controlled environments, parallel to some of the conclusions we've drawn from our models — specifically, that the future success of a book, movie, celebrity, or Web site is strongly influenced by these types of feedback effects, and hence may to some extent be inherently unpredictable.

Closer Relationships between Power Laws and Information Cascades? Considerations of this sort suggest an important question for further research: understanding the relationship between power laws and information cascades at a deeper level. When we looked at information cascades, we saw how a population in which people were aware of earlier decisions made between two alternatives (e.g. accepting an idea or rejecting it) could end up in a cascade, even if each person is making an optimal decision given what they've observed. Our copying model for power laws draws on the intuition behind this model, but it differs in

several respects. First, a model for popularity should include choices among many possible options (e.g. all possible Web pages), rather than just two options. Second, the copying model involves a set of people who engage in very limited observation of the population: when you create a new Web page, the model assumes you consult the decision of just one other randomly selected person. And third, the copying model is based on the idea that later people imitate the decisions of earlier people, but it doesn't derive this imitation from a more fundamental model of rational decision-making.

The first two of these differences simply reflect the specifics of the problem being modeled here — the way in which popularity evolves over time. But it would be very interesting to overcome the third of these differences, constructing a copying-style model to produce power laws on top of a base model of individual decision-making. Such an approach could shed further insight into the mechanisms behind rich-get-richer dynamics, and provide a picture of popularity as arising from competing information cascades whose intensities vary according to the power laws that we observe in real systems.

18.5 The Long Tail

The distribution of popularity can have important business consequences, particularly in the media industry. In particular, let's imagine a media company with a large inventory — a giant retailer of books or music, for example — and consider the following question: are most sales being generated by a small set of items that are enormously popular, or by a much larger population of items that are each individually less popular? In the former case, the company is basing its success on selling "hits" — a small number of blockbusters that create huge revenues. In the latter case, the company is basing its success on a multitude of "niche products," each of which appeals to a small segment of the audience.

In a widely-read 2004 article entitled "The Long Tail," Chris Anderson argued that Internet-based distribution and other factors were driving the media and entertainment industries toward a world in which the latter alternative would be dominant, with a "long tail" of obscure products driving the bulk of audience interest [13]. As he wrote, "You can find everything out there on the Long Tail. There's the back catalog, older albums still fondly remembered by longtime fans or rediscovered by new ones. There are live tracks, B-sides, remixes, even (gasp) covers. There are niches by the thousands, genre within genre within genre: Imagine an entire Tower Records devoted to '80s hair bands or ambient dub."

Although sales data indicates that the trends are in fact somewhat complex [146], this tension between hits and niche products makes for a compelling organizing framework. It also accords with the fundamental models of companies like Amazon or Netflix, where the ability to carry huge inventories — without the restrictions imposed by physical stores — makes it feasible to sell an astronomical diversity of products even when very few of them

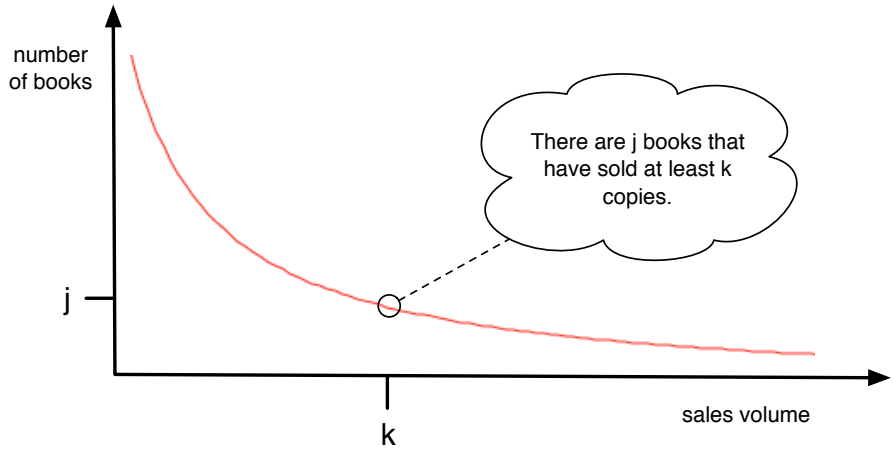

Figure 18.3: The distribution of popularity: how many items have sold at least k copies?

generate much volume on their own. And ultimately, quantifying the importance of the Long Tail comes down to an analysis of power laws.

Visualizing the Long Tail. The first thing to notice, when we compare this discussion of the Long Tail to our earlier analysis of power laws, is that in some sense we're now viewing things out the opposite end of the telescope. Initially, we started from a baseline in which we expected to see Gaussian distributions and tight concentration around the average, and then we observed that the number of highly popular items was much higher than this baseline would suggest. Now, on the other hand, we're starting from a very different default view of the world — a sort of stereotype of the media business in which only blockbusters matter — and we're observing that the total sales volume of *unpopular* items, taken together, is really very significant. In terms of the plot in Figure 18.2, this new view is focusing on the upper-left part of the plot, whereas before we were primarily focused on the lower-right.

Once you recognize that this contrast is going on, it's not hard to reconcile the two views [4]. First, let's modify our original definition of the popularity curve slightly, in a way that doesn't fundamentally change what we're measuring. Rather than asking

As a function of k, what fraction of items have popularity exactly k?

let's instead ask

As a function of k, what number of items have popularity at least k?

Notice that we've changed two things: "fraction" to "number" (a completely inconsequential change), and "exactly k" to "at least k". This second change modifies the function we're

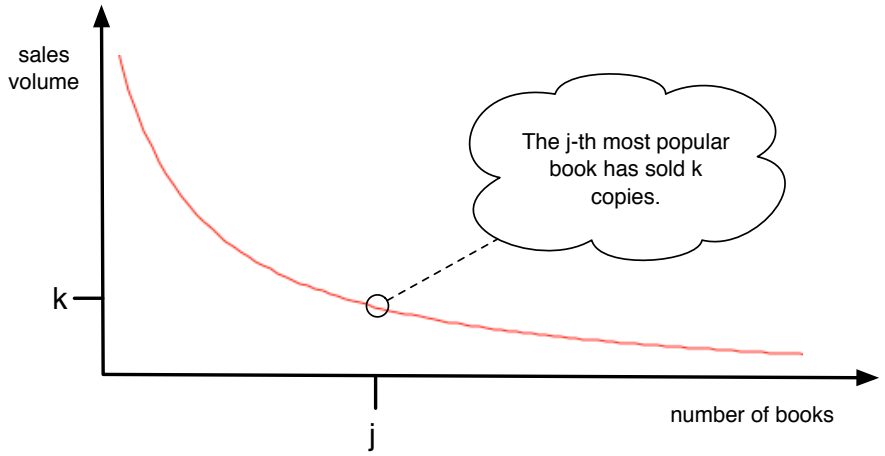

Figure 18.4: The distribution of popularity: how many copies of the j^{th} most popular item have been sold?

considering; but while we won't go through the derivation here, it's possible to show that if the original function was a power-law, then this new one is too. We show a schematic plot of this new function in Figure 18.3; if we're talking about the popularity of some item like books, then a point (k, j) on this curve means, by definition, "There are j books that have sold at least k copies."

So far, this is still the conceptual view from the previous section: as we follow the x-axis of the curve to the right, we're essentially asking, "As you look at larger and larger sales volumes, how few books do you find?" To capture the discussions of the Long Tail more directly, we want to be asking the following question as we follow the x-axis to the right: "As you look at less and less popular items, what sales volumes do you see?"

If we think about it, this simply involves switching the two axes. That is, suppose that we plot exactly the same curve, but we interchange the roles of the x- and y-axes, as shown in Figure 18.4. Interpreting this new curve literally from its definition, a point (j, k) on the curve says, "The j^{th} most popular book has sold k copies." This is exactly what we want: we order the books by "sales rank," and then we look at the popularity of books as we move out to larger and larger sales ranks — into the niche products.[1] And the characteristic shape of this curve, tailing off slowly downward to the right, is the visual basis for the term "Long Tail."

One can now easily discuss trends in sales volume, and their consequences, in terms of the curve in Figure 18.4. Essentially, the area under the curve from some point j outward is

[1] Our notion of "sales rank" simply reflects the sorted, decreasing order of all items by sales volume. When the term "sales rank" is used by on-line retailers such as Amazon, it tends to be a more complex measure that incorporates other factors as well.

the total volume of sales generated by all items of sales-rank j and higher; and so a concrete version of the hits-vs.-niche question, for a particular set of products, is whether there is significantly more area under the left part of this curve (hits) or the right (niche products). And the debate over trends toward niche products becomes a question of whether this curve is changing shape over time, adding more area under the right at the expense of the left.

It is worth noting that curves of the type in Figure 18.4 — with the axes ordered so that the variable on the x-axis is rank rather than popularity — have a long history. They are often called *Zipf plots* after the linguist George Kingsley Zipf, who produced such curves for a number of human activities [423]. Most famously, he identified the empirical principle known as *Zipf's Law*, that the frequency of the j^{th} most common word in English (or most other widespread human languages) is proportional to $1/j$. Thus, perhaps not surprisingly, the debate within the media industry about curves like this echoes earlier fascination in other areas.

18.6 The Effect of Search Tools and Recommendation Systems

We conclude by briefly discussing a further question that has been growing in importance as people consider popularity and its distribution: are Internet search tools making the rich-get-richer dynamics of popularity more extreme or less extreme? What is interesting is that there are two compelling but juxtaposed sides to this question, and its ultimate resolution will likely involve decisions about how individuals and corporations design and deploy future generations of search tools.

On one side of this question, we've seen that a model in which people copy links from uniformly random Web pages already gives an advantage to popular pages. But once people are using search engines such as Google to find pages, then even the choice of what to copy from becomes highly skewed: as we've seen, Google is using popularity measures to rank Web pages, and the highly-ranked pages are in turn the main ones that users see in order to formulate their own decisions about linking. A similar argument can be made for other media in which a handful of the most popular items have the potential to crowd out all others. In simple models, this kind of feedback can accentuate rich-get-richer dynamics, producing even more inequality in popularity [103].

There are other forces at work, however. To begin with, users type a very wide range of queries into Google, and so there isn't a single list of "top pages on Google" — rather, by getting results on relatively obscure queries, users are being led to pages that they are likely never to have discovered through browsing alone. Search tools used in this style, targeted more closely to users' specific interests, can in fact provide ways around universally popular pages, enabling people to find unpopular items more easily, and potentially counteracting

the rich-get-richer dynamics. Here too, simple mathematical models have demonstrated how such effects can work [165].

This latter view also forms an important part of Anderson's Long-Tail argument: in order to make money from a giant inventory of niche products, a company crucially needs for its customers to be aware of these products, and to have some reasonable way to explore them [13]. Viewed in this light, the types of *recommendation systems* that companies like Amazon and Netflix have popularized can be seen as integral to their business strategies: they are essentially search tools designed to expose people to items that may not be generally popular, but which match user interests as inferred from their history of past purchases.

Ultimately, the design of search tools is an example of a kind of higher-order feedback effect: by causing people to process their available options in one way or another, we can reduce rich-get-richer effects, or amplify them, or potentially steer them in different directions altogether. These are among the subtle consequences that take place when we inject sophisticated information systems into what is an already complex social system.

18.7 Advanced Material: Analysis of Rich-Get-Richer Processes

In Section 18.3, we described a simple model of a growing directed network based on copying — or equivalently, based on rich-get-richer dynamics. We claimed there that the fraction of nodes with k in-links is distributed approximately according to a power law $1/k^c$, where c depends on the behavior of nodes in the model. Here we provide a heuristic argument that analyzes the behavior of the model to indicate why the power law arises, and in fact goes further to show how the power-law exponent c is related to more basic features of the model. The analysis is based on the simple differential equation governing exponential growth that one sees in introductory calculus.

First, let's reprise the description of the model from Section 18.3, as follows.

(1) Pages are created in order, and named $1, 2, 3, \ldots, N$.

(2) When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule (which is controlled by a single number p between 0 and 1).

   (a) With probability p, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to this page i.

   (b) With probability $1 - p$, page j chooses a page ℓ with probability proportional to ℓ's current number of in-links, and creates a link to ℓ.

   (c) This describes the creation of a single link from page j; one can repeat this process to create multiple, independently generated links from page j. (However, to keep

things simple, we will suppose that each page creates just one outbound link.)

Note that here we are using the rich-get-richer version of step (2b), rather the original copying version. Recall that the two formulations are equivalent; but for our purposes, it will be easier to analyze the phrasing of the model as we have it here.

At one level, we now have a purely probabilistic question: we have specified a randomized process that runs for N steps (as the N pages are created one at a time), and we can simply determine the expected number of pages with k in-links at the end of the process. (Or analyze the distribution of this quantity.) While several groups of researchers have performed this analysis [68], it is a bit more intricate than what we can feasibly cover here. Instead, we describe how an approximation to the model allows for a much simpler calculation that gives the correct value for the exponent c in the power law. This approximate analysis was the first one to be carried out [42], and it thus provided heuristic evidence for the power-law effect that was then verified by more rigorous analysis of the full probabilistic model.

A deterministic approximation of the rich-get-richer process. Before describing the approximation to the model, let's discuss some simple properties of the original probabilistic model itself. First, the number of in-links to a node j at a time step $t \geq j$ is a random variable $X_j(t)$. Let's observe two facts about $X_j(t)$.

(a) *The initial condition.* Since node j starts with no in-links when it is first created at time j, we know that $X_j(j) = 0$.

(b) *The expected change to X_j over time.* Node j gains an in-link in step $t+1$ if and only if the link from the newly created node $t+1$ points to it. What is the probability that this happens? With probability p, node $t+1$ links to an earlier node chosen uniformly at random, and with probability $1-p$ it links to an earlier node with probability proportional to the node's current number of in-links. In the former case, node $t+1$ links to node j with probability $1/t$. For the latter case, we observe that at the moment node $t+1$ is created, the total number of links in the network is t (one out of each prior node), and of these, $X_j(t)$ point to node j. Thus, in the latter case, node $t+1$ links to node j with probability $X_j(t)/t$. Therefore, the overall probability that node $t+1$ links to node j is

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}.$$

The basic plan in building an approximation to the model is to analyze a different, closely analogous, but simpler rich-get-richer process, in which it is correspondingly easier to discover the power law. Again, this does not directly imply that the original model behaves the same way, but the similarities between the two models offer evidence that can then be verified by further analysis of the original model.

The central idea in formulating the simpler model is to make it *deterministic* — that is, a model in which there are no probabilities, but in which everything instead evolves in a fixed way over time, like an idealized physical system that behaves according to some "equations of motion" starting from a set of initial conditions. To do this, we have time run continuously from 0 to N (rather than in the discrete steps $1, 2, 3, \ldots$), and we approximate $X_j(t)$ — the number of in-links of node j — by a continuous function of time $x_j(t)$. We characterize the function x_j by two properties that seek to approximate the initial conditions and expected change over time that we described above for $X_j(t)$. The two properties of the function x_j are the following.

(a) *The initial condition.* Recall that $X_j(j) = 0$. We define $x_j(j) = 0$ as well.

(b) *The growth equation.* Recall that when node $t + 1$ arrives, the number of in-links to node j increases with probability

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$$

In the determinstic approximation provided by the function x_j, we model this rate of growth by the differential equation

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}. \tag{18.1}$$

Using differential equations, we have thus specified the behavior of x_j, our deterministic approximation to the number of in-links to node j over time. Essentially, rather than dealing with random variables $X_j(t)$ that move in small probabilistic "jumps" at discrete points in time, we get to work with a quantity x_j that grows completely smoothly over time, at a rate tuned to match the expected changes in the corresponding random variables.

We now explore the consequences of the differential equation defining x_j; this leads quickly to the kind of power-law distribution we want.

Solving the deterministic approximation. We begin by solving the differential equation (18.1) governing x_j. For notational simplicity, let's write $q = 1-p$, so that the differential equation becomes

$$\frac{dx_j}{dt} = \frac{p + qx_j}{t}.$$

Dividing both sides by $p + qx_j$, we get

$$\frac{1}{p + qx_j}\frac{dx_j}{dt} = \frac{1}{t}.$$

Integrating both sides

$$\int \frac{1}{p + qx_j}\frac{dx_j}{dt}dt = \int \frac{1}{t}dt,$$

we get

$$\ln(p + qx_j) = q \ln t + c$$

for a constant c. Exponentiating, and writing $A = e^c$, we get

$$p + qx_j = At^q$$

and hence

$$x_j(t) = \frac{1}{q}\left(At^q - p\right). \tag{18.2}$$

Now, we can determine the value of the constant A by using the initial condition $x_j(j) = 0$. This condition gives us the equation

$$0 = x_j(j) = \frac{1}{q}\left(Aj^q - p\right),$$

and hence $A = p/j^q$. Plugging this value for A into Equation (18.2), we get

$$x_j(t) = \frac{1}{q}\left(\frac{p}{j^q} \cdot t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]. \tag{18.3}$$

Identifying a power law in the deterministic approximation. Equation (18.3) is a significant intermediate step in the analysis, since it gives us a closed-form expression for how each x_j grows over time. Now we want to use this to ask the following question: For a given value of k, and a time t, what fraction of all nodes have at least k in-links at time t? Since x_j approximates the number of in-links of node j, the analogue to this question that we consider in our simplified model is: For a given value of k, and a time t, what fraction of all functions x_j satisfy $x_j(t) \geq k$?

Using Equation (18.3), this corresponds to the inequality

$$x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k,$$

or, re-writing this in terms of j,

$$j \leq t\left[\frac{q}{p} \cdot k + 1\right]^{-1/q}.$$

Out of all the functions $x_1, x_2, \ldots, x_t$ at time t, the fraction of values j that satisfy this is simply

$$\frac{1}{t} \cdot t\left[\frac{q}{p} \cdot k + 1\right]^{-1/q} = \left[\frac{q}{p} \cdot k + 1\right]^{-1/q}. \tag{18.4}$$

We can already see the power law taking shape here: since p and q are constants, the expression inside brackets on the right-hand-side is proportional to k, and so the fraction of x_j that are at least k is proportional to $k^{-1/q}$.

For the final step, note that this has so far been about the fraction of nodes $F(k)$ with *at least* k in-links. But from this, we can directly approximate the fraction of nodes $f(k)$ with *exactly* k in-links simply by taking the derivative — in other words, approximating $f(k)$ by $-dF/dk$. Differentiating the expression in Equation (18.4), we get

$$\frac{1}{q}\frac{q}{p}\left[\frac{q}{p}\cdot k+1\right]^{-1-1/q}.$$

In other words, the deterministic model predicts that the fraction of nodes with k in-links is proportional to $k^{-(1+1/q)}$ — a power law with exponent

$$1+\frac{1}{q}=1+\frac{1}{1-p}.$$

Subsequent analysis of the original probabilistic model showed that, with high probability over the random formation of links, the fraction of nodes with k in-links is indeed proportional to $k^{-(1+1/(1-p))}$ [68]. The heuristic argument supplied by the deterministic approximation to the model thus provides a simple way to see where this power-law exponent $1+1/(1-p)$ comes from.

The behavior of this exponent also makes sense intuitively as we vary p. When p is close to 1, link formation is mainly based on uniform random choices, and so the role of rich-get-richer dynamics is muted. Correspondingly, the power-law exponent tends to infinity, showing that nodes with very large numbers of in-links become increasingly rare. On the other hand, when p is close to 0, the growth of the network is strongly governed by rich-get-richer behavior, and the power-law exponent decreases toward 2, allowing for many nodes with very large numbers of in-links. The fact that 2 is a natural limit for the exponent as rich-get-richer dynamics become stronger also provides a nice way to think about the fact that many power-law exponents in real networks (such as for the number of in-links to a Web page) tend to be slightly above 2.

A final appealing feature of this deterministic analysis is that it is very malleable — it can be easily modified to cover extensions of the model, and this has been the subject of considerable further research [10].

18.8 Exercises

1. Consider an on-line news site, such as `cnn.com` or `nytimes.com`, which consists of a front page with links to many different articles. The people who operate such sites generally track the popularity of the various articles that get posted, asking questions like the ones that we've seen in this chapter: "As a function of k, what fraction of all articles have been viewed by k people?" Let's call this the *popularity distribution* of the articles.

Now suppose that the operators of such a news site are considering changing the front page, so that next to each link is a counter showing how many people have clicked on the link. (E.g., next to each link it will say something like, "30,480 people have viewed this story," with the number getting updated over time.)

First, what effect do you think this change will have on the behavior of people using the site? Second, do you expect that adding this feature will cause the popularity distribution of the articles to follow a power-law distribution more closely or less closely, compared to the version of the site before these counters were added? Give an explanation for your answer.

2. When we covered power laws in Chapter 18, we discussed a number of cases in which power laws arise, generally reflecting some notion of "popularity" or a close analogue. Consider, for example, the fraction of news articles each day that are read by k people: if $f(k)$ represents this fraction as a function of k, then $f(k)$ approximately follows a power-law distribution of the form $f(k) \approx k^{-c}$ for some exponent c.

   Let's think about this example in more detail, and in particular consider the following questions. What mechanisms for providing news to the public will tend to accentuate this power-law effect, causing the most widely-read articles to be even more widely-read? What mechanisms will tend to diminish the power-law effect, more evenly balancing readership across more and less widely-read articles? Give an explanation for your answer.

   This is an open-ended question, in the sense that the range of possible correct answers is quite broad; also, it is fine for your answer to be informally stated, provided the explanation is clear.

3. Suppose that some researchers studying educational institutions decide to collect data to address the following two questions.

   (a) As a function of k, what fraction of Cornell classes have k students enrolled?

   (b) As a function of k, what fraction of 3rd-grade elementary school classrooms in New York State have k pupils?

   Which one of these would you expect to more closely follow a power-law distribution as a function of k? Give a brief explanation for your answer, using some of the ideas about power-law distributions developed in Chapter 18.